DOCUMENT RESUME

ED 124 581 TM 005 337

AUTHOR Crambert, Albert C.
TITLE Use of Mastery Cutoff Scores in Criterion-Referenced Measurement.
PUB DATE [Apr 76]
NOTE 17p.; Paper presented at the Annual Meeting of the American EDucational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Cutting Scores; Decision Making; *Literature Reviews; Measurement Techniques; *Scoring; *Standards; Test Interpretation

ABSTRACT

There are two major aspects to the cutoff score issue in criterion-referenced (CR) measurement: whether a cutoff score is actually needed for a CR test, and the method for setting the cutoff score if one is used. There are many factors to be considered in determining a cutoff score. While cutoff scores are very often set arbitrarily (e.g., 80%), there have been many methods suggested to improve the quality of judgment or to utilize quantitative approaches of varying degrees of complexity and precision; these methods are reviewed in this paper. Although more research is needed to confirm the value of these suggested methods, several appear promising. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

USE OF MASTERY CUTOFF SCORES IN CRITERION-REFERENCED MEASUREMENT[1]

Albert C. Crambert

U.S. Office of Education, Philadelphia, Pa.[2]

ED124581

One of the issues which distinguishes criterion-referenced (CR) measurement fron norm-referenced (NR) measurement is that of setting a "cutoff" or "mastery" score denoting the level of minimum acceptable performance on the segment of instruction covered by the test. This issue seldom arises in NR measurement, because NR interpretations ordinarily are made on a relative rather than an absolute basis. However, the cutoff score issue is one of the key points of controversy among the various conceptualizations of CR measurement. There are two major aspects to the cutoff score issue: whether a cutoff score is, in fact, actually needed, and the method which should be used to establish a cutoff score if one is to be used.

There is one position in the literature on CR measurement which holds that a cutoff score is not considered necessary or relevant; in this view, CR measurement does not necessarily imply making absolute judgments. This position is well expressed by Nitko, who takes the following view:

---

[1]Presented at the annual meeting of the American Educational Research Association, San Francisco, April 23, 1976.

[2]This paper was written by the author in his private capacity. No official support or endorsement by the U.S. Office of Education is intended or should be inferred.

TM005 337

> absolute interpretations can be estremely dangerous . . . Nothing in the nature of CR testing implies that anyone necessarily meet a given standard of competence, only that such levels of competency be defined in terms of performance (Nitko, 1970, p. 39).

While he does not rule out the usefulness of a cutoff score in some situations, Nitko contends that the concept of CR testing does not necessarily require making a value judgment abcut whether flawless performance is possible.

Nitko prefers an empirical, decision-oriented approach, where a cutoff score is not set on an a priori basis but only on an empirical basis in answer to a question such as,"What level of performance is required at one point in the instructional sequence in order to maximize success at the next point in the sequence?" There is no inherent reason why this point could not differ among individuals and in different circumstances.

This point of view as expressed by Nitko is unusual. The more typical view in the CR literature is that the setting of a cutoff score is an inherent requirement of CR measurement which is often reflected even in the definition of a CR test. Depending on the writer, the requirement for a cutoff score may be considered as fundamental to the development of the CR test (e.g., Ivens, 1970; Jackson, 1970) or it may be considered as essential in making an interpretation of the score (e.g., Fremer, 1972).

Usually, the task of a CR test is to place an examinee in one of two categories--master or nonmaster--with a minimum number of classification errors. Most of the methods to be described are concerned with this kind of dichotomous

decision. However, it also would be very reasonable to consider models which divide examinees into three or more categories (Harris, 1974).

The establishment of a cutoff (or mastery) score involves some difficult methodological problems. However, it also involves resolving some basic conceptual issues as well. Skager (Note 1) notes that "How to define the nature of any performance that would indicate mastery of a somain of content remains a major conceptual problem (p. 14)." At a practical level, the fact that item difficulties for CR tests, as for other tests, can be easily influenced presents a danger of making incorrect assumptions that any given score on a CR test represents an accurate judgment as to mastery or nonmastery of an objective (Klein & Kosecoff, 1973). Stanley and Hopkins (1972) have demonstrated that, as might be expected, items of widely different difficulty can be written to fit the same instructional objective. Therefore, criteria such as 90 percent are arbitrary and nearly meaningless in the absence of a definitive external reference point. Furthermore, an arbitrary cutoff level, such as 80%, implicitly assumes that all items are of equal importance (Kifer & Bramble, 1974), a quite unreasonable assumption.

Setting a minimum performance level prior to instruction is particularly appropriate when mastery is essential to the subsequent attainment of other important objectives (Sullivan, 1969). In other words, the setting of a cutoff level for an objective should properly rest upon a determination that the attainment of the objective has instructional significance.

The Standards for Educational and Psychological Tests (American Psychological Association, et al., 1974) also require that a rationale be provided for the selection of a cutoff score used in test interpretation. To accept this goal, however, does not determine how it can be achieved. What is desired is to minimize the number of incorrect classification decisions, but neither classical (NR) procedures nor item sampling approaches are very effective in individual decisionmaking (Haladyna, 1975).

Like many issues in CR measurement, the issue of setting a mastery standard has a long history. Monroe (1917), for example, discussed the issue at some length and concluded that a standard must meet two conditions: that it be reasonable and that it be "efficient." A reasonable standard was defined by Monroe as one which realistically can be attained by students, and an "efficient" standard was defined as one which represents a level of performance which equips students for meeting present and future demands.

The level at which a cutoff score should be set will vary depending upon the cruciality of the objective; for very important objectives, the appropriate cutoff level may be quite high. Two other important factors should be considered in establishing performance standards: the difficulty level of the instructional content (insofar as this can be determined independently of actual learner performance), and the amount of instructional time available relative to the instructional material to be covered.

Minimum performance standards may be established in one of two ways: by setting a cutoff score which must be attained

by each individual learner, or by specifying a group standard in terms of the percentage of students in the class or other target group who will be expected to attain the individual cutoff score if the instruction is successful. The latter approach is much less common.

An important point in the setting of the criterion level was made by Kriewall and Hirsch (1969), who pointed out that "setting a higher error criterion does not of itself improve the proficiency found in those examinees classified as masters (p. 8)." Similarly, poorer-performing individuals do not become even less proficient by being designated as non-masters. That is, the distribution of actual proficiency is independent of the imposition of proficiency standards (Gardner, 1962). Simply by moving the standard, the proportion of masters or nonmasters can be changed without having any effect upon the distribution of actual performance. Mastery standards cannot be set independent of the performance of the individuals involved; the level of performance which may be required for mastery must be realistic in terms of the prevailing levels of competence (Garvin, 1971).

There rarely is a clear basis upon which to establish a cutoff score in educational situations. In the absence of other evidence, the cutoff score is most commonly set on some subjective basis relying on informed judgment. Gronlund (1973), for example, has offered a step-by-step, trial-and-error procedure in which an initial arbitrary standard is then adjusted upward or downward on the basis of experience and judgment. Gronlund proceeds from the suggestion of Block (1971) that,

although 100 percent mastery might seem to be the ideal, 80 to 85 percent correct is a more realistic standard.

Gronlund's procedure is workable, but it is only a general guide to exercising informed judgment. The procedure does not unequivocably determine the size of the adjustments or the level of the final standard. Like Nitko, Gronlund (1973) concluded that the ultimate question is, "What level of mastery is necessary in order to learn effectively at the next stage of instruction (p. 13)?"

Such subjectivity in setting cutoff levels represents a serious shortcoming of CR measurement to some writers such as Ebel (1973). Ebel has also pointed out that relative standards, as in NR measurement, which are derived from the average performances of groups of examinees are more stable than absolute standards based on the judgments of individual instructors.

Millman (1973) is among those critical of routinely using a single percentage standard in all content domains and for all individuals. He suggests five approaches which might be used to standardize and refine the application of judgment in establishing standards of achievement.

1. Standards established on the basis of the actual past performance of typical persons, so that some predetermined percentage of persons will pass. This approach is applicable when only a fixed number or proportion can be permitted to "pass," and it probably resembles typical NR practice more than it does criterion-referencing. Block (1971) and Klein (1972) likewise suggest transferring existing grading standards

set under non-mastery conditions to the mastery situation.

2. Standards established on the basis of informed judgment, on an item-by-item basis, as to how important it is that each item on a test be answered correctly. This would preferably be done by a panel of informed judges working under standardized procedures. Nedelsky's (1954) procedure for assigning grades, described below, may be considered as one variation of this approach, and Aulls and Pearson's (Note 2) method of "quantifying intuitions" is another. Still another variation of this approach, for the test as a whole, has been suggested by Kriewall (1973).

3. Standards established on the basis of educational consequences in terms of future learning. Higher cutoff scores may be required for fundamental or prerequisite learnings. Weighted regression equations or expectancy tables may be useful in this approach.

4. Standards established on the basis of psychological or financial costs, so far as these can be determined. A high cutoff may be justified when the cost of false advances is high.

5. Standards established with allowance made for measurement error due to pure guessing, or for the effects of non-representativeness of the item sample. This latter approach is actually a suggestion for final refinement of whatever cutoff score may be derived by one of the other approaches.

While any of the approaches discussed by Millman would help to improve the process of setting cutoff scores, it must be recognized that most of them are relatively crude and dependent upon unsubstantiated judgment and knowledge of actual

performance. Other more systematic approaches to the problem have been proposed in attempts to objectify the process of setting a cutoff level. Some of these more systematic approaches are still entirely judgmental, and others are statistical in nature.

Nedelsky (1954) addressed the problem of determining the cutoff level indirectly, through the assignment of grades. In Nedelsky's approach, the "Minimum Passing Score" from which other grades are scaled is derived solely on the basis of pooled judgment without reference to the distribution of actual obtained scores. The system is exact and unambiguous in its application, but it is not intuitively appealing. Nedelsky's procedure has not come into significant use.

Another more systematic, but still somewhat subjective, view of establishing a cutoff score is offered by Fremer (1972). Fremer suggested five methods by which a measure can be given CR meaning, urging that more than one of the methods should be applied in any particular situation. Most of Fremer's methods are independent of knowledge of observed scores. Fremer's suggested methods have not been developed in great depth, and they are not unequivocal in their results. They are:

1. The use of nontest information to set a minimal performance standard, such as by determining, a priori, that only the top 10% will "pass."

2. Teacher judgments of individual test items to estimate the proportion of a group of "barely passing" students who would answer the item correctly. The judgments of a number of knowledgeable raters are then averaged to obtain the minimum

criterion cutoff score for the total instrument. This approach was followed by the Educational Testing Service to develop instruments for the statewide Michigan Assessment Program, with apparently satisfactory results.

3. Teacher judgments regarding which students are performing at minimum competency levels, either through global judgments or a detailed analysis of student performance.

4. Development of supplemental work sample tests as criteria against which to validate the CR measure. This method closely resembles the traditional predictive validity approach. Ward (Note 3) used an approach of this sort to "validate" CR tests used in teacher training.

5. Development of what Fremer calls "stand alone" work sample tests. These are instruments constructed to serve as direct measures of performance on objectives which are considered so important that they should be measured directly even though efficient indirect measurement might be possible.

The preceding approaches mostly represent different ways to arrive at a refined judgment as to the cutoff score. A number of quantitative approaches to the cutoff score problem, of varying degrees of sophistication, have also been proposed. However, even the supposedly objective approaches to be described still require judgment at some key points and often involve normative comparisons (Messick, Note 4).

One of the most straightforward of these is that of Millman (Note 5) who uses the binomial model to derive tables representing an exact solution to the question of how many items are needed to attain a given level of precision of measurement for a

particular proficiency level. The establishment of a proficiency level to which the procedure would apply is still left as basically a subjective process. This approach is appropriate to the degree to which the items which comprise the test can be assumed to be a ramdom sample of a defined universe of items. An unknown degree of error would be introduced by using the Millman approach for CR tests in which the items cannot be assumed to represent a defined universe. Millman's tables can be used to answer either of two types of mathematically parallel, reciprocal questions: (a) for individual assessment, how many items are needed on a test? and (b) for program evaluation, how many students should be tested? The tables provide a means to determine the proportion of misclassifications to be expected when a test of a given length is administered with a given passing score. An elaboration of Millman's binomial procedure has been proposed by Novick and Lewis (1974) in which prior probabilities are used in a Bayesian model to relate observed test scores to true level of functioning.

A second quantitative approach is proposed by Block (1972). His approach is based on utilizing students' future learning as a criterion for determining the level of proficiency (mastery standard) which students must have attained at intermediate stages of instruction. It is, in effect, an operational answer to the type of question posed by Gronlund, Millman, and Nitko: "What level of mastery is needed in order to succeed at the next level of instruction?" Any of several statistical techniques may be used for this purpose. That level of interim performance which yields the greatest estimated future learning

is selected as the mastery standard. Block's system appears to be sufficiently comprehensive to form a basis for designing an instructional strategy. However, considerable subjectivity still remains in his system.

A third statistical approach is suggested by Kriewall (1972), who suggests using an item sampling model to derive an exact probability that a CR test will provide a false negative or false positive result. Kriewall's procedure does not require an assumption of equal item difficulty or an assumption of a true dichotomy between mastery vs. nonmastery status (Millman, 1973) but test statistics are affected by test length. The procedure does not determine the mastery level to be used, but it provides a basis for evaluating the outcome of any particular cutoff score chosen.

A promising, relatively simple but comprehensive statistical approach to the cutoff score problem is offered by Emrick (1971). Emrick's "skill-mastery test model" seeks to establish mastery (cutoff) scores which are optimized in terms of relative decision error costs and relative item error probabilities. It combines item and student information to produce probability statements regarding skill-mastery status. Skill mastery is treated as an all-or-none variable. Emrick's model has considerable statistical elegance. However, some of the assumptions underlying the model such as homogeneity, equal item difficulties, and equal item intercorrelation, frequently are not tenable in practice (Millman, 1973), and the consequences of violating these assumptions are not known. The Emrick model also remains subjective at some key points,

but nevertheless appears to offer two advantages: (a) by carefully isolating and defining the subjective judgments, it permits them to be made more accurately; and (b) for a given set of input values, the formula yields the single optimum cutoff score.

An approach somewhat similar to Emrick's, but which attempts to avoid some of these difficulties, is offered by Besel (1973). Besel's "mastery learning test model" also differs from Emrick's in another important respect, by using an independent estimate of the proportion of students in a comparison group which have achieved an objective (which may be looked at as normative data) as a check upon the accuracy of individual prediction. This use of group estimates of prior probabilities resulted in significantly improved stability of mastery learning parameters and improved individual predictions. Although Besel's model overcomes some of the limiting assumptions of Emrick's approach, it is mathematically complex and seems doubtful that Besel's approach offers a worthwhile advantage over Emrick's simpler model.

A still more complex approach to the cutoff score issue is the "decision-theoretic" approach (Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, Note 6). This is a Bayesian procedure which allows the use of prior and collateral information, and also incorporates the cost of misclassifications. The procedure deliberately introduces the decision-maker's values into the decision process. The decision-theoretic approach will accomodate a three-category decision system. However, like some other approaches which have been discussed,

the decision-theoretic approach does not actually establish a cutoff score. Rather, it starts with an arbitrarily-set cutoff score and then analyzes the consequences of using that cutoff.

Among all of these many techniques for establishing cutoff scores on CR tests which have been discussed, and still others which have been suggested in the literature, one can be found to meet almost any situation in which a cutoff score must be established. Nearly any of the methods will be an improvement over the too-common practice of arbitrarily setting a cutoff score. Although little research has been conducted to confirm the value of the various methods, several appear very promising. For classroom use by teachers not highly skilled in measurement, several of the commonsense methods suggested by Millman or Fremer should be useful, and, for larger applications, Emrick's "skill-mastery test model" particularily appears to warrant wider use.

Reference Notes

1. Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly and interpretation. Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973.

2. Aulls, M. W., & Pearson, P. D. The search for standards of performance for criterion-referenced measures of reading achievement. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, April 1974.

3. Ward, B. A. Establishing the standard of performance. Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973.

4. Messick, S. The standard problem: Meaning and values in measurement and evaluation. Paper presented at the meeting of the American Psychological Association, New Orleans, August 1974.

5. Millman, J. Tables needed for determining number of items needed on domain-referenced tests and number of students to be tested (Technical Paper No. 6). Los Angeles: Instructional Objectives Exchange, 1972.

6. Swainathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Amherst, Mass.: University of Massachusetts, 1974.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D. C.: American Psychological Association, 1974.

Besel, R. Using group performance to interpret individual responses to criterion-referenced tests. Los Alamitos, Calif.: SWRL Educational Research and Development, 1973. (Professional Paper 25.) (ERIC Document Reproduction Service No. ED 076 658.)

Block, J. H. Introduction to mastery learning: Theory and practice. In J. H. Block (Ed.), Mastery learning: Theory and practice. New York: Holt, Rinehart & Winston, 1971.

Block, J. H. Student evaluation: Toward the setting of mastery performance standards. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972. (ERIC Document Reproduction Service No. ED 065 605.) (a)

Block, J. H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-191. (b)

Ebel, R. L. Evaluation and educational objectives. Journal of Educational Measurement, 1973, 10, 273-279.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Fremer, J. Criterion-referenced interpretations of survey achievement tests. Princeton, N. J.: Educational Testing Service, 1972. (Report No. ETS-TDM-72-1.) (ERIC Document Reproduction Service No. ED 065 533.)

Gardner, E. F. Normative standard scores. Educational and Psychological Measurement, 1962, 22, 7-14.

Garvin, A. D. The applicability of criterion-referenced measurement by content area and level. In W. J. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Gronlund, N. E. Preparing criterion-referenced tests for classroom instruction. New York: Macmillan, 1973.

Haladyna, T. An analysis of two procedures for decision-making when using domain-referenced tests. Paper presented at the meeting of the National Council on Measurement in Education, Washington, D. C., April 1975. (ERIC Document Reproduction Service No. ED 104 957.)

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation, No. 3). Los Angeles: UCLA Center for the Study of Evaluation, 1974.

Ivens, S. H. An investigation of item analysis, reliability, and validity in relation to criterion-referenced tests. (Doctoral dissertation, Florida State University.) Dissertation Abstracts International, 1971, 32, 4548-A. (University Microfilms No. 71-7036)

Jackson, R. Developing criterion-referenced tests. Princeton, N. J.: Educational Testing Service, 1970. (ERIC-TM Report 1.) (ERIC Document Reproduction Service No. ED 041 052.)

Kifer, E., & Bramble, W. The calibration of a criterion-referenced test. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1974. (ERIC Document Reproduction Service No. ED 091 434.)

Klein, S. P. Ongoing evaluation of educational programs. Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972. (ERIC Document Reproduction Service No. ED 069 725.)

Klein, S. P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. Princeton, N. J.: Educational Testing Service, 1973. (ERIC-TM Report 26.) (ERIC Document Reproduction Service No. ED 083 284.)

Kriewall, T. E. Remarks addressed to AERA regarding criterion-referenced tests. Illinois School Research, 1973, 9(2), 19-21.

Kriewall, T. E., & Hirsch, E. The development and interpretation of criterion-referenced tests. Paper presented at the meeting of the American Educational Research Association, Los Angeles, 1969. (ERIC Document Reproduction Service No. ED 042 815.)

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Monroe, W. S. Educational tests and measurements. Boston: Houghton Mifflin, 1917.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Nitko, A. J. Criterion-referenced testing in the context of instruction. In Testing in turmoil: A conference on problems and issues in educational measurement. New York: Educational Records Bureau, 1970.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement (CSE Monograph Series in Evaluation, No. 3). Los Angeles: UCLA Center for the Study of Evaluation, 1974.

Stanley, J. C., & Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Sullivan, H. J. Objectives, evaluation, and improved learner achievement. In W. J. Popham (Ed.), Instructional objectives. Chicago: Rand McNally, 1969.